Out[1]:

Out[2]: Click here to toggle on/off the Python code.

PHYS 10792: Introduction to Data Science

2018-2019 Academic Year

The University of Manchester

Course instructors: Rene Breton (http://www.renebreton.org) - Twitter @BretonRene (https://twitter.com/BretonRene)
Marco Gersabeck (http://www.hep.manchester.ac.uk/u/gersabec) - Twitter @MarcoGersabeck (https://twitter.com/MarcoGersabeck)

Chapter 8

Syllabus

1. Probabilities and interpretations
2. Probability distributions
3. Parameter estimation
4. Maximum likelihood + extended maximum likelihood
5. Least square, chi2, correlations
6. Monte Carlo basics
7. Probability
8. **Hypothesis testing**
9. Confidence level
10. Goodness of fit tests
11. Limit setting
12. Introduction to multivariate analysis techniques

Topics

8 Hypothesis testing

8.1 Decision making

- 8.1.1 Introductory examples
- 8.1.2 Hypotheses
- 8.1.3 Alternative hypotheses
- 8.1.4 Types of errors
- 8.1.5 Significance and power

8.2 Practical examples

- 8.2.1 Hypotheses and alternative hypotheses
- 8.2.2 Interpreting experiments: Null hypothesis

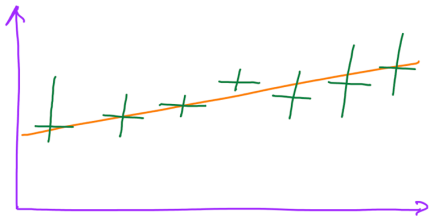
8 Hypothesis testing

8.1 Decision making

Hypothesis testing is essntially decision making. Any test boils down to a yes-or-no answer. This means it is fundamentally different to asking what the measured value of a given observable is as that can in principle take an infinite number of values. Rather, hypothesis taking can evaluate whether the measured value of an observable represents a certain behaviour or not.

Example

Rather than asking what the value of the slope of a fit like the one shown below is, we can ask whether the slope indicates that the linear function is increasing. This is not yet a very precise question and we will pin down the various loose ends as we progress through this lecture.



8.1.1 Introductory examples

In general we need the following to conduct a hypothesis test:

- The assertion that some hypothesis is true,
- A numerical test that is to be applied to data, and
- A hypothesis that is accepted or rejected depending on the outcome of the test.

There are several examples that will all feature in the remainder of this course:

- The interpretation of experiments
- Goodness of fit tests
- Two-sample problems
- Analyses of several samples

Any hypothesis test will not be infallible. We need to choose a level of confidence at which to take the decision. Going back to our example of the linear fit, this level of confidence will be linked to the measured value and its uncertainty. We will discuss how to quantify this in the following.

8.1.2 Hypotheses

Hypotheses are statements that are either true or false. We distinguish simple and composite hypotheses.

Simple hypotheses define the probability distribution function completely. Example hypotheses are:

- These data are drawn from a Poisson distribution of mean 3.4.
- The new treatment has identical effects to the old.

Composite hypotheses combine several probability distribution functions. Example hypotheses are:

- These data are drawn from a Poisson distribution of mean greater than 4.
- The new treatment is an improvement on the old.

Referring back to the components of a hypothesis test, a simple hypothesis corresponds to a numerical test featuring a single equality that is being evaluated, while an inequality would be the numerical test of a composite hypothesis.

8.1.3 Alternative hypotheses

In hypothesis tests we often compare to alternative hypothses. These can take many different forms; examples of alternatives to the first hypothesis (Poisson with mean 3.4) are:

- These data are drawn from a Poisson distribution with another given mean, e.g. 4.5
- These data are drawn from a Poisson distribution with any random mean other than 3.4
- These data are not drawn from a Poisson distribution. This needs to be specified further in order to yield a numerical test.

In general, it is crucial to distinguish between one-tailed directional and two-tailed non-directional tests. A two-tailed test refers to the comparison of a test outcome to a value where we don't care of whether the outcome is less than or greater than the value. In the directional test the sign of the difference between test outcome and comparison value is of importance.

8.1.4 Type I/II errors

The outcome of a hypothesis test is the acceptance or rejection of the hypothesis based on the numerical test. However, it may be that the decision taken does not reflect whether or not the hypothesis is actually true. The two cases where there is a mismatch are called Type I and Type II error according to the following pattern:

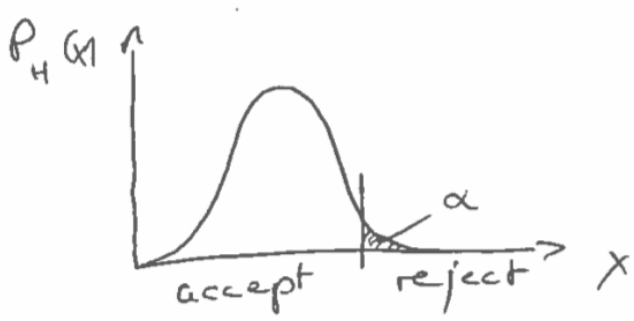
Hypothesis \ Decision	accept	reject
true	:)	Type I error
false	Type II error	:)

8.1.5 Significance and Power

Significance

Type I errors are inevitable and the rate at which they occur is called significance. We know the probabbility ditribution function of the hypothesis, $P_H(x)$, for the case that the test involves the measurement of a quantity x. In our previous example, $P_H(x)$ would be a Poisson distribution with mean 3.4.

The probability distribution function is then divided into a rejection and an acceptance region.



The decision is then taken depending on where the measured value of x falls.

The significance, α, is the integral of the probability distribution of the hypothesis over the rejection region:

The significance, α, is the integral of the probability distribution of the hypothesis over the rejection region:

$$\alpha = \int_{Reject} P_H(x)dx.$$

Typically, we want α to be small, e.g. 1% or 5%. In reality, we often need to work with the inequality

$$\alpha \geq \int_{Reject} P_H(x)dx,$$

as we may have a range of possible $P_H(x)$ (composite hypothesis) or a discrete distribution for which α cannot be reached exactly. For composite hypotheses we would use the $P_H(x)$ that maximises α. All other incarnations of the composite hypothesis would therefore result in a smaller integral and be equally accepted.

Power

Considering the alternative hypothesis, we can define the integral of the probability distribution of the alternative hypothesis over the acceptance region as

$$\beta = \int_{Accept} P_A(x)dx,$$

or, by integrating of the rejection region as above, we get

$$1 - \beta = \int_{Reject} P_A(x)dx,$$

where $1 - \beta$ is called the power of the test.

On the choice of the accept and reject regions

The sketches above show the accept region to the left of the reject region. This is just one possible arrangement and the optimal way depends on the situation at hand. The general guiding principle is that we want

- to minimise the significance, and
- to maximise the power.

Therefore, if the alternative hypothesis has a probability distribution to the right of the hypothesis under test the arrangement above makes sense. If the alternative hypothesis were to lie on the left, then accept and reject regions should be swappend. Finally, it is also possible to have a central accept region with reject regions to either side. We will see examples for that when discussing confidence intervals.

Example

Let us consider the following hypothesis:

A data point is drawn from a Poisson distribution with mean less or equal to 5.

We want to test this with 5% significance.

Hence, we need to evaluate

$$\alpha \geq \int_{Reject} P_H(x)dx,$$

which turns into

$$0.05 \geq \int_{Reject} Poisson(x; \lambda = 5)dx.$$

The Poisson distribution is dicrete, so we are looking for the limit, n, of the sum that satisfies

$$0.05 \geq \sum_{x=n+1}^{\infty} Poisson(x; \lambda = 5),$$

or, to avoid an infinite sum,

$$1 - 0.05 = 0.95 < \int_{Accept} Poisson(x; \lambda = 5)dx = \sum_{x=0}^{n} Poisson(x; \lambda = 5).$$

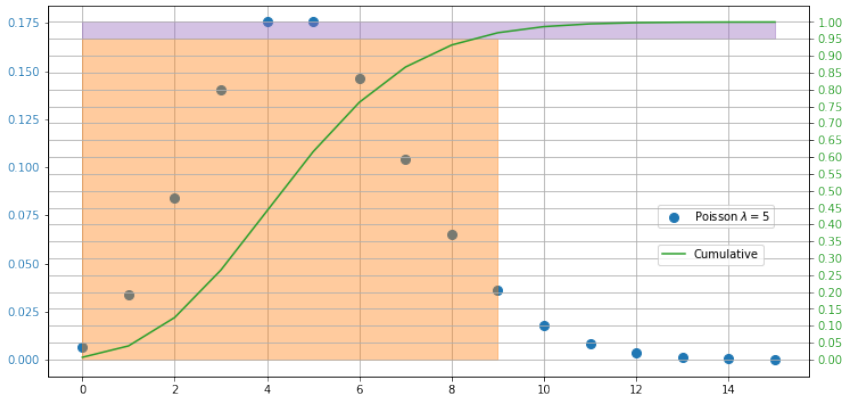
Here, n has been chosen as the limit of the acceptance region.

For $n = 9$ the cumulative sum is 0.968, which is the first n for which it exceeds 0.95.

A greater value of n would satisfy these inequalities as well, but we want to choose the n that leaves us closest to the target value of 0.95.

Any smaller mean than 5 satisfies the equations above as well with $n = 9$, which is why we chose to work with with $\lambda = 5$.

Therefore, any observed value of 9 or smaller would have the hypothesis accepted.



Quiz

What do we want from significance and power?

$$\alpha = \int_{Reject} P_H(x)dx; \qquad 1-\beta = \int_{Reject} P_A(x)dx.$$

menti.com (https://www.menti.com) with key 88 22 34.

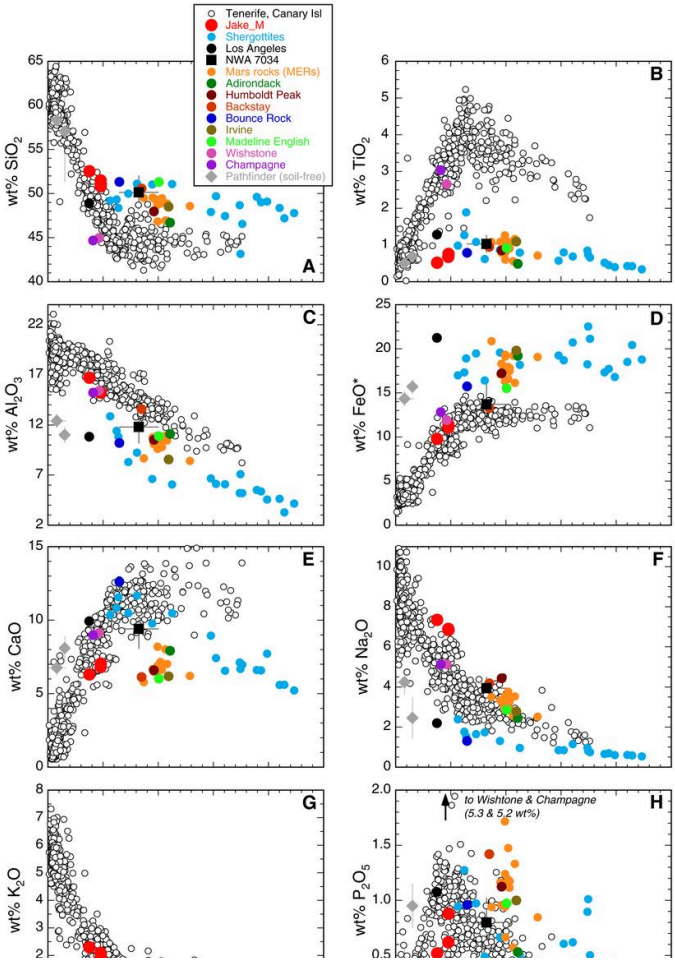
8.2 Practical examples

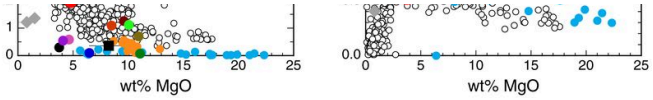
8.2.1 Hypotheses and alternative hypotheses

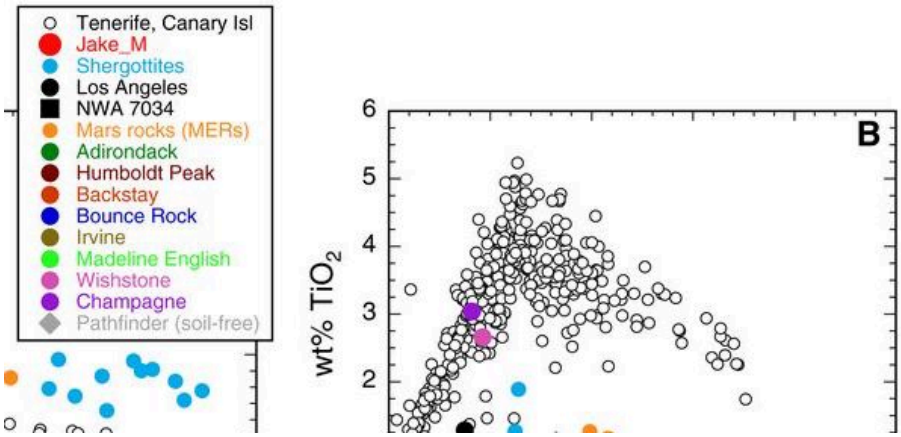
Example: Identifying Mars rocks

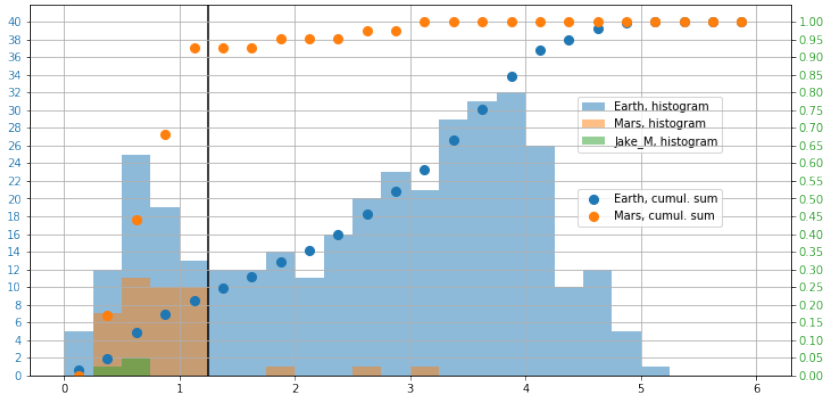
There has been no return mission to Mars, so all Martian rocks on earth came through being ejected from the surface of Mars by a meterorite impact and then having found their way to us. Their identity is compared to in-situ measurements by probes on the surface of Mars.

The following plot is from E.M. Stolper et al., Science 6153 (2013) 1239463.









Defining the acceptance region for the alternative hypothesis (earth-like) as $> 1.25\%$, we would wrongly reject about 7% of Martian rocks, i.e. $\alpha = 7\%$, and we would wrongly accept about 21% of earth rock samples, hence $\beta = 21\%$, which means the power is 79%.

The sample analysed in this paper are the Jake_M rocks, for which we would reject the alternative hypothesis of being of earth origin in this test. A power of 79% may not be very satisfactory in this context and the correlation shown in the plots illustrate that a combined analysis of all available input would be much preferrable. We will get back to this topic in Chapter 11.

8.2.2 Interpreting experiments: null hypothesis

Measurements can have different goals; some are designed to measure a certain quantity, e.g. the rate at which a process occurs, while others aim to make a discovery or test a particular theory prediction. The former is rather straightforward in terms of statistics as it just requires knowledge of the resolution of the experiment (and any possible systematic uncertainties). Statistical treatment is at best required for repeated executions of the measurements, which is sufficiently trivial.

When testing the validity of a theory or aiming to make a discovery, we find ourselves in a situation where we are testing a hypothesis. However, we can only ever reject a hypothesis with great confidence, but not accept it. This is because any given measurement will yield a result for an observable that is a random variable distributed according to the probability distribution of the hypothesis. Any measured value that falls in the bulk of the probability distribution does not indicate a strong increase or decrease in the support for the hypothesis, whereas a value measured to be in the extreme tails of the distribution allows to reject the hypothesis at great confidence.

In short: for any theory we want to test, we have to formulate the opposite hypothesis and aim to falsify this. This hypothesis is called the null hypothesis, H_0.

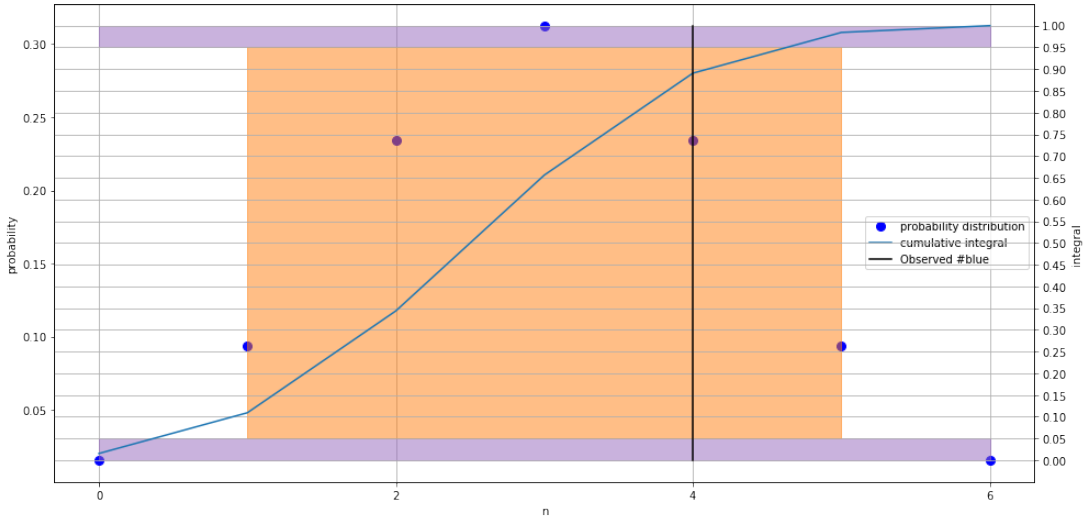
Example: Red versus Blue

Let's generate some data: menti.com (https://www.menti.com) with key 88 22 34.

We want to test whether there's a bias in a preference for red or blue among the students in the lecture. Hence, we need to formulate the oppostite as our null hypothesis: there is no bias.

If there is no bias, the probability for either colour is $1/2$. Let's test our hypothesis to the 10% level. In this context we have a Binomial distribution

if there is no bias, the probability for either colour is 1/2. Let's test our hypothesis to the 10% level. In this context we have a Binomial distribution.



Purple bands indicate probability range where hypothesis would be rejected. The orange box indicates the corresponding region of n_{blue} (or n_{red}) where hypothesis is *not* rejected.

Example: Binomial statistics

Let's look at a different example of binomial statistics now. In our very first lecture we encountered a medical diagnostic tool and medical applications are a very typical use-case for Binomial statistics.

Let's assume that we're looking at a treatment that is tested on 100 patients. 60% of these are expected to be cured spontaneously within a week. Let's work at the 5% significance level.

The null hypothesis is that the probability of a patient being cured within a week is $P \leq 0.6$; hence, we expect to get 60 cures or less under this hypothesis with a standard deviation of

$$\sigma = \sqrt{100 \times 0.4 \times 0.6} = \sqrt{24} = 4.9.$$

Using a normal approximation, we know that a one-tailed distribution reaches 5% at 1.64σ. Hence, we set the decision point at $60 + 1.64 \times 4.9 = 68.03$ cures. We have to round this up to 69, which we can then use as the threshold to reject the hypothesis that the number of cures agrees with the expected level of spontaneous cures.

Example: Poisson statistics

In many counting experiments we will have a situation where we want to identify whether a significant excess exists above a certain level of background.

Let's join a group of bird watchers who make a head count of a bird population once per year and who want to know whether the population grew since the previous year with a significance of 1%. The 2017 count was 132, followed by 160 in 2018.

We want to test the hypothesis that the 2018 count is in agreement with the same distribution as observed in 2017, i.e. a Poisson distribution with mean 132. The corresponding uncertainty would be $\sqrt{132} = 11.5$. Hence, we see an increase of $(160 - 132)/11.5\ \sigma = 2.4\sigma$. This corresponds to a significance of 0.7% for a one-tailed distribution.